



ALFRED P. SLOAN
FOUNDATION



The Dataverse Project - Data sharing, Reproducibility, Research, Development and the Community

CNSTAT Expert meeting on Guidance on Data Sharing
for NIA Longitudinal Studies
May 10, 2021

Ana Trisovic, Harvard University
on behalf of the Dataverse Project team



- A free and open-source software platform to archive, share, and cite research data
 - Focus on data sharing and making data available
- Provides data repository software that can be installed at institutions
 - Supports research communities for entire countries (NO, NL)
- Developed at Harvard's Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community
 - 122 contributors to the software



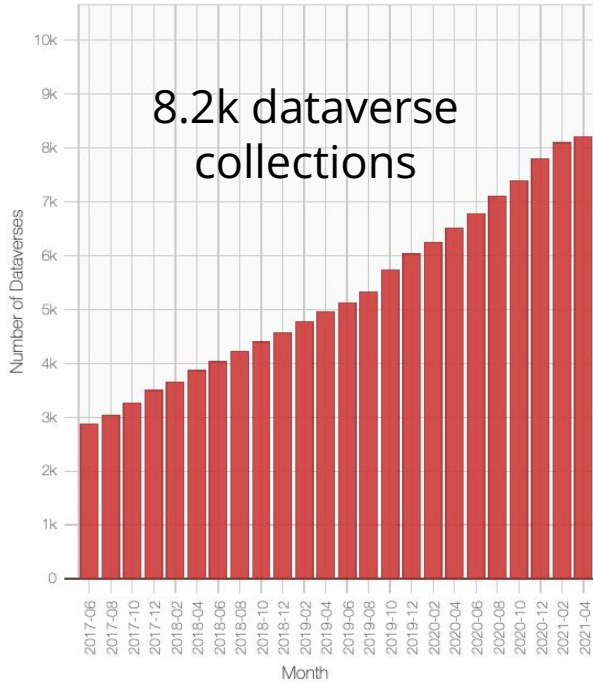
- A free and open-source software platform to archive, share, and cite research data
 - Focus on data sharing and making data available
- Provides data repository software that can be installed at institutions
 - Supports research communities for entire countries (NO, NL)
- Developed at Harvard's Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community
 - 122 contributors to the software

69 institutions around the globe run Dataverse installations as their official data repository

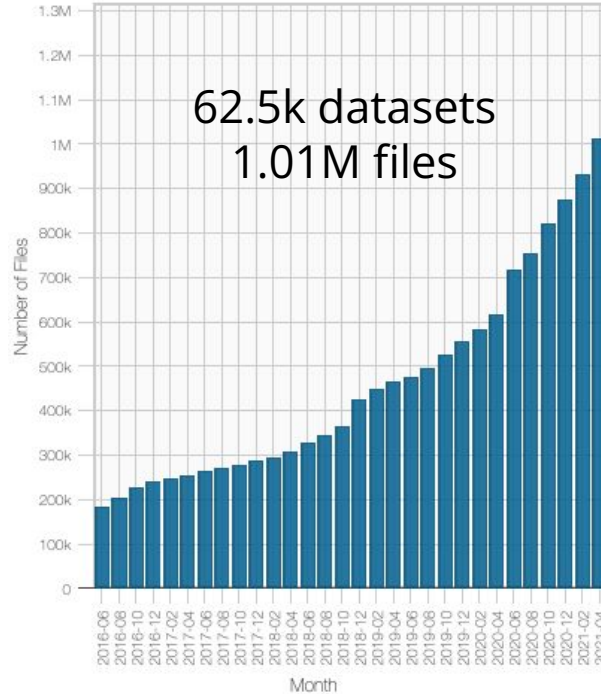


Metrics

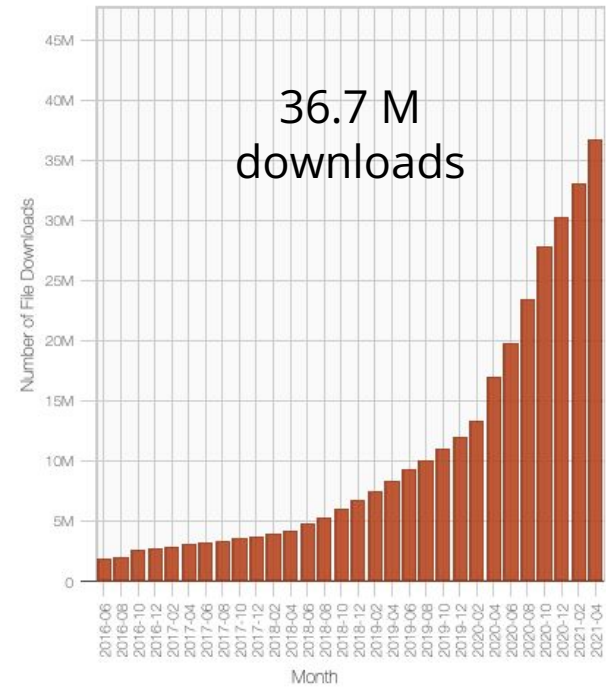
Total Dataverse Collections



Total Files



Total File Downloads



<https://dataverse.org/metrics> (29 Dataverse installations)

Presentation agenda

1. Data sharing at Dataverse repositories
2. Documentation and guidelines
3. Reproducibility and reuse
4. Open questions, research and development

Data sharing

Data sharing

- Stand-alone or institutional account for depositing data
 - Dataverse installation account
 - GitHub, Google, ORCID, University credentials
- Two approaches for data sharing:
 - Via UI in the web browser
 - Via API (command line or Dataverse Software clients)

The screenshot displays the Harvard Dataverse web interface for creating a new dataset. At the top, the Harvard Dataverse logo is visible, along with navigation links for 'Add Data', 'Search', 'About', 'User Guide', and 'Support'. The main section is titled 'Host Dataverse' and includes a text input field containing 'Harvard Dataverse'. Below this, a note states: 'Changing the host dataverse will clear any fields you may have entered data into.' A warning message reads: '*Asterisks indicate required fields'. The 'Citation Metadata' section is expanded, showing several required fields: 'Title' (text input), 'Author' (Name, Affiliation, Identifier Scheme, Identifier), 'Contact' (Name, Affiliation, E-mail), 'Description' (Text area), 'Subject' (dropdown), 'Keyword' (Term, Vocabulary), and 'Related Publication' (Citation). Each field has a plus sign icon to its right, indicating it can be added or modified. The 'Author' and 'Contact' sections are pre-filled with the name 'Trisovic, Ana' and affiliation 'Harvard University'. The 'E-mail' field is pre-filled with 'anatrisovic@fas.harvard.edu'. The 'Date' field is pre-filled with 'YYYY-MM-DD'. The 'Vocabulary URL' field has a placeholder text: 'Enter full URL, starting with http://'. The 'Citation' field is currently empty.

Dataverse Software Clients

A client is an application that connects to and communicates with the remote Dataverse installation to transfer or manage data through the API.



pyDataverse



dataverse-client-javascript



dataverse-client- java



dataverse-client-python



dataverse-client-r

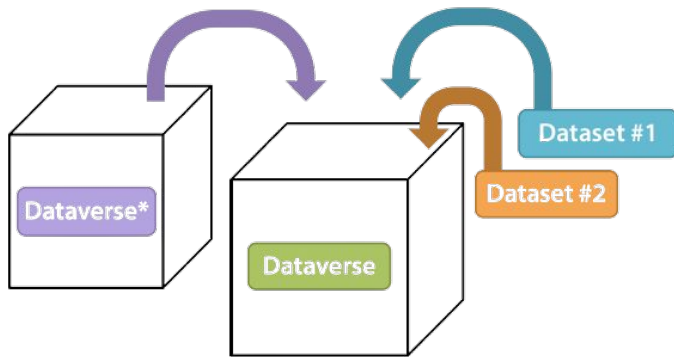


dataverse_api

All clients are open-source and available on GitHub.

Dataverse collection and datasets

- A dataverse collection is a collection of datasets and/or other collections
- Individuals, institutes or journals may have own dataverse collections



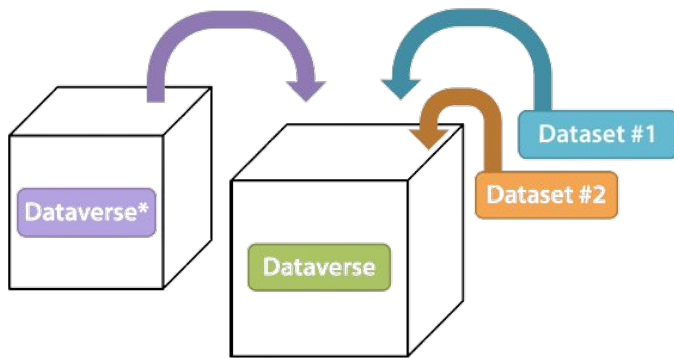
Container for your **Datasets** and/or **Dataverses***

The screenshot shows the website for the Henry A. Murray Research Archive at Harvard University. The header includes the name 'Henry A. Murray Research Archive at Harvard University' and a portrait of Henry A. Murray. Below the header, the text reads 'Murray Research Archive Dataverse (Harvard University) Home'. A navigation link for 'Harvard Dataverse >' is visible. On the right, there are links for 'Contact' and 'Share'. The main content area states: 'The Henry A. Murray Research Archive is the endowed, permanent repository for quantitative and qualitative research data at the Institute for Quantitative Social Science. Our collection comprises over 125 terabytes of data, audio, and video.' It also mentions that all 'restricted' data require an 'application for the use of data' form and provides a link to the website for more information. At the bottom, there is a carousel of dataset categories: 'Diversity Datasets: Race, Ethnicity, Sexual Orientation, Religion Dataverse', 'Early Head Start Research and Evaluation Project Dataverse', 'Economic Theory and Demography Dataverse', and 'Education Dataverse'.

<https://murray.harvard.edu>

Dataverse collection and datasets

- A dataverse collection is a collection of datasets and/or other collections
- Individuals, institutes or journals may have own dataverse collections



Container for your **Datasets** and/or **Dataverses***

The screenshot shows the Henry A. Murray Research Archive website. The header includes the name 'Henry A. Murray Research Archive at Harvard University' and a 'Home' link. Below the header, there is a navigation bar for 'Murray Research Archive Dataverse (Harvard University)'. The main content area shows a search for 'Texas Adoption Project' on the 'Harvard Dataverse' platform. The search results show one result: 'Texas Adoption Project' by Joseph Horn and John Loehlin, published in 2010. The result includes a link to the DOI and a brief description of the study.

Replication dataset

- Replication dataset - a bundle of data, code and other files needed to reproduce a published study

HARVARD Dataverse Add Data Search About

POLITICAL ANALYSIS

Political Analysis Dataverse (Cambridge University Press)

HARVARD Dataverse Add Data Search About User Guide Support

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org

Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse

The *American Journal of Political Science* is committed to significant advances in knowledge and understanding of citizenship, gov politics, and to the public value of political science research. To find out more about our data integrity policies, please visit our web

Replication Data for: How Political Parties Shape Public Opinion in the Real World

Version 2.0



Bisgaard, Martin; Rune Slothuus, 2020, "Replication Data for: How Political Parties Shape Public Opinion in the Real World", <https://doi.org/10.7910/DVN/Z5BTCQ>, Harvard Dataverse, V2, UNF-6:YTyX+kjbsSZUNEND/3GGg== [fileUNF]

[Cite Dataset -](#)

[Learn about Data Citation Standards.](#)

Access Dataset -
Contact Owner Share

Dataset Metrics
1,092 Downloads

Description

How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan elites and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real world. We present a rare quasi-experimental panel study of how citizens responded when their political party suddenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinions changed immediately and substantially when their party switched its policy position—even when the new position went against citizens' previously held views. These findings advance the current, largely experimental literature on partisan elite influence. (2020-03-26)

Subject

Social sciences

Keyword

Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey

Related Publication

Bisgaard, Martin, and Rune Slothuus. [date], "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* Forthcoming. <http://ajps.org/>

Notes

This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the Center for Open Science.



Files Metadata Terms Versions

Search this dataset...

Filter by
File Type: All - Access: All -

Sort -

1 to 10 of 25 Files Download -

	build_data.R R Syntax - 12.1 KB Published Jun 29, 2020 56 Downloads MD5: a94...597		
	codebook_ess.pdf Adobe PDF - 508.8 KB Published Jun 29, 2020 46 Downloads		

Replication dataset

- Replication dataset - a bundle of data, code and other files needed to reproduce a published study

Replication Data for: How Political Parties Shape Public Opinion in the Real World

Version 2.0

Bisgaard, Martin; Rune Slothuus, 2020, "Replication Data for: How Political Parties Shape Public Opinion in the Real World", <https://doi.org/10.7910/DVN/Z5BTCQ>, Harvard Dataverse, V2, UNF-6:YTyX+kjbsSZUNEND/3GGg== [fileUNF]

[Cite Dataset](#)

[Learn about Data Citation Standards.](#)

[Access Dataset](#)

[Contact Owner](#) [Share](#)

[Dataset Metrics](#)

1,092 Downloads

Dataset metrics

Description

How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan elites and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real world. We present a rare quasi-experimental panel study of how citizens responded when their political party suddenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinion immediately and substantially when their party switched its policy position—even if their position went against citizens' previously held views. These findings advance the experimental literature on partisan elite influence. (2020-03-26)

Subject

Social Sciences

Keyword

Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey

Related Publication

Bisgaard, Martin, and Rune Slothuus. [date]. "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* Forthcoming. <http://ajps.org/>

Notes

This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the [Center for Open Science](#).



Center for Open Science Badges

[Files](#) [Metadata](#) [Terms](#) [Versions](#)

Search this dataset...

Filter by
File Type: All - Access: All -

Sort

1 to 10 of 25 Files

Download

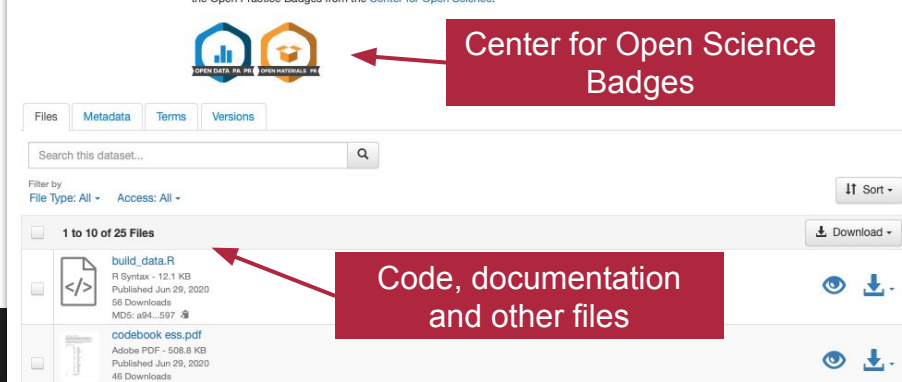


[build_data.R](#)
R Syntax - 12.1 KB
Published Jun 29, 2020
56 Downloads
MD5: a04...597

Code, documentation and other files



[codebook_ess.pdf](#)
Adobe PDF - 508.8 KB
Published Jun 29, 2020
46 Downloads



Support for multiple metadata standards

Rich support for
metadata standards in
**human and machine
readable formats.**

Export Metadata ▾

- Dublin Core
- DDI
- DataCite
- DDI HTML Codebook
- JSON
- OAI_ORE
- OpenAIRE
- Schema.org JSON-LD

Files

Metadata

Terms

Versions

Citation Metadata ^

Dataset Persistent ID ?

doi:10.7910/DV

Previous Dataset Persistent ID ?

hdl:1902.1/000

Publication Date ?

2009-03-05

Title ?

Early Head Start Research and Evaluation Project, 1996 - 2001

Other ID ?

00097

Author ?

Administration for Children and Families (U.S. Department of Health & Human Services)

Description ?

This study page contains cataloging and documentation files (only) related to the *Early Head Start* data archived in the Murray Research Archive Dataverse.

The purpose of this study was to assess the impact of early head start programs in response to the 1994 Head Start reauthorization which established a special initiative for services to families with infants and toddlers. The study was a program evaluation with 1500 families in Early Head Start programs and 1500 in a control group with no program participation.

Tabular files and variable metadata

ClimateRegressionData_150327.tab

This file is part of "Replication Data for: "Climate Amenities, Climate Change, and American Quality of Life" Journal of the Association of Environmental and Resource Economists 3, no. 1 (March 2016): 205-246."

Version 1.0

File Citation

Albouy, David, Graf, Walter, Kellogg, Ryan, and Wolff, Hendrik, 2018, "ClimateRegressionData_150327.tab", *Replication Data for: "Climate Amenities, Climate Change, and American Quality of Life" Journal of the Association of Environmental and Resource Economists 3, no. 1 (March 2016): 205-246.*, <https://doi.org/10.7910/DVN/QCE1XY/BNJLIA>, Harvard Dataverse, V1, UNF:6:CBIOoHJrG5/T6i+XjwBVwg== [fileUNF]

[Cite Data File](#)

[Learn about Data Citation Standards.](#)

Dataset Citation

Albouy, David, Graf, Walter, Kellogg, Ryan, and Wolff, Hendrik, 2018, "Replication Data for: "Climate Amenities, Climate Change, and American Quality of Life" Journal of the Association of Environmental and Resource Economists 3, no. 1 (March 2016): 205-246.", <https://doi.org/10.7910/DVN/QCE1XY>, Harvard Dataverse, V1, UNF:6:CBIOoHJrG5/T6i+XjwBVwg== [fileUNF]

[Cite Dataset](#)

[Learn about Data Citation Standards.](#)

[Preview](#)

[Metadata](#)

[Versions](#)

[Explore on View Data](#)

	mssname	Wage_orig	Wage	Price	QOL_orig	QOL_25_1	
47	Phoenix-Mesa, AZ	-0.0607973	0.014620701	-0.13433664	-0.013151387	0.027213141	-0
48	Phoenix-Mesa, AZ	-0.031059736	0.014049928	-0.20824265	-0.050257135	-0.021000927	-0
49	Phoenix-Mesa, AZ	0.13126019	0.025910802	0.1713764	-0.0038431883	-0.088416085	0.1
50	Phoenix-Mesa, AZ	0.083860636	0.024022257	0.08576128	-0.009372421	-0.062420316	0.1
51	Phoenix-Mesa, AZ	0.08740685	0.024419874	0.05565375	-0.020567477	-0.07349341	0.1

Variable metadata from tabular data files

```
var ..
  @ID: v17547221
  @name: gdpgrowthrate
  @intrvl: contin
  location
    @fileid: f3002597
  sumstat 106.27980041503906
    @type: max
  sumstat .
    @type: mode
  sumstat 5791.0
    @type: vald
  sumstat 6.563398040445294
    @type: stdev
  sumstat 3.8374838829040527
    @type: medn
  sumstat -51.030860900878906
    @type: min
  sumstat 2309.0
    @type: invd
  sumstat 3.6461661697240384
    @type: mean
  varformat
    @type: numeric
    @formatname: float
  notes UNF:6:IWNGf1eqM40D8xQJ
    @subject: Universal Numeric f
    @level: variable
    @type: VDC:UNF
```

Tabular files and variable metadata

Access File ▾

- Download Options ▾
- Original File Format (Stata Binary)
- Tab-Delimited
- RData Format
- Variable Metadata
- Data File Citation
- Explore Options ▾
- Data Explorer**
- View Data

Replication Data for: Politics by Number: Indicators as Social Pressure in International Relations kelley_simmons_ajps_2014_replication.tab

Kelley, Judith; Simmons, Beth, 2018, "Replication Data for: Politics by Number: Indicators as Social Pressure in International Relations" [fileUNF]

Q x 7 Results Download ▾

ID	Name	Label
17547274	crim1date	
17547149	crimdate1	
17547168	crimdate2	
17547244	crimdate3	
17547265	unctocdate	
17547283	protocoldate	
17547148	tier_date	

Variable search

Chart View Table View

Variable crimdate1:

Values Categories

Summary Statistics

Cases	N
	5977
	2123
Maximum	2012
Minimum	1967
	2006.4731470637444
	2007
	6.89046957527742

155 variables

Tabular files and variable metadata

Replication Data for: Politics by Number: Indicators as Social Pressure in International Relations
kelley_simmons_ajps_2014_replication.tab

Kelley, Judith; Simmons, Beth, 2018, "Replication Data for: Politics by Number: Indicators as Social Pressure in International Relations", <https://doi.org/10.7910/DAN/DK0001>
UNF:6:IEYJGdv4VpPallzdj0qJNg== [fileUNF]

Q date x 7 Results Download

ID	Name	Label
17547274	crim1date	
17547149	crimdate1	
17547168	crimdate2	
17547244	crimdate3	
17547265	unctocdate	
17547283	protocoldate	
17547148	tier_date	

Chart View Table View

Variable crimdate1:

Values	Categories
Summary Statistics	
Cases	N
	5977
	2123
Maximum	2012
Minimum	1967
	2006.4731470637444
	2007
	6.89046957527742

Extensive variable metadata (descriptive statistics) automatically derived from tabular data file in DDI format

Data sharing

Summary

- Dataverse data repository software
 - Versatile support for data sharing via UI and the clients
 - Extensive support in the repository for data, variables and metadata
-

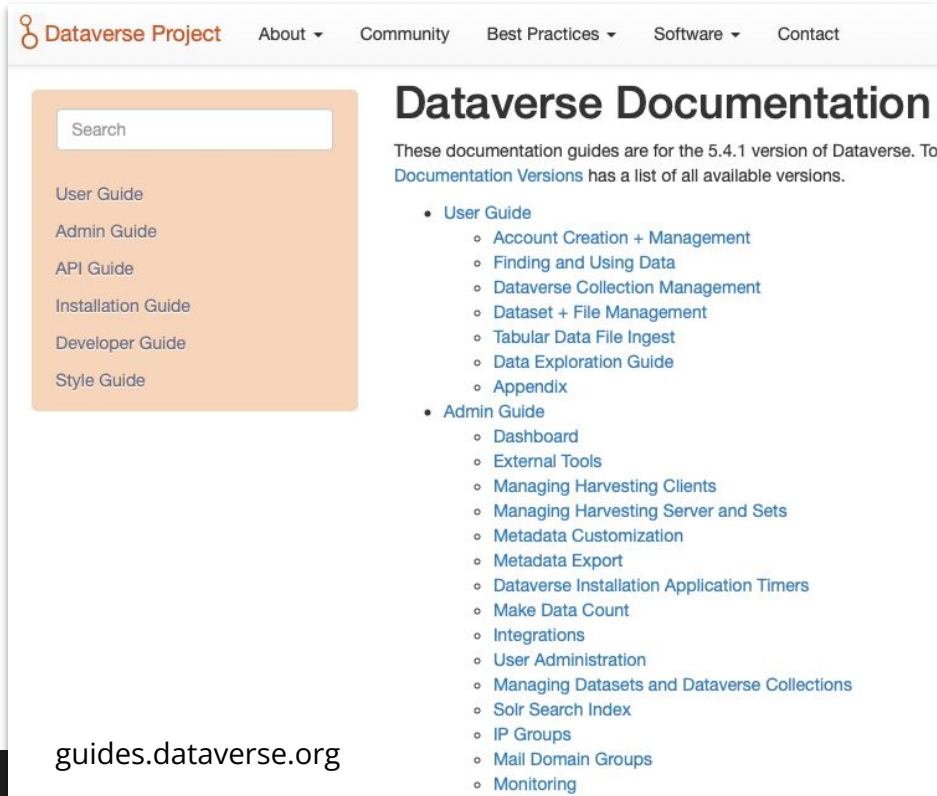
Documentation and guidelines

Documentation and guidelines

The Dataverse Project team maintains an extensive set of guidelines for repository managers, developers and users.

Dataverse Community meetings

- projects.iq.harvard.edu/dcm2021
- June 15-17, 2021



The screenshot shows the Dataverse Project documentation website. The header includes the logo and navigation links: About, Community, Best Practices, Software, and Contact. The main content area is titled "Dataverse Documentation" and includes a search bar and a list of guides. The guides are categorized into User Guide, Admin Guide, and API Guide, with sub-items for each. The URL guides.dataverse.org is displayed at the bottom.

Dataverse Project About Community Best Practices Software Contact

Dataverse Documentation

These documentation guides are for the 5.4.1 version of Dataverse. To [Documentation Versions](#) has a list of all available versions.

- **User Guide**
 - [Account Creation + Management](#)
 - [Finding and Using Data](#)
 - [Dataverse Collection Management](#)
 - [Dataset + File Management](#)
 - [Tabular Data File Ingest](#)
 - [Data Exploration Guide](#)
 - [Appendix](#)
- **Admin Guide**
 - [Dashboard](#)
 - [External Tools](#)
 - [Managing Harvesting Clients](#)
 - [Managing Harvesting Server and Sets](#)
 - [Metadata Customization](#)
 - [Metadata Export](#)
 - [Dataverse Installation Application Timers](#)
 - [Make Data Count](#)
 - [Integrations](#)
 - [User Administration](#)
 - [Managing Datasets and Dataverse Collections](#)
 - [Solr Search Index](#)
 - [IP Groups](#)
 - [Mail Domain Groups](#)
 - [Monitoring](#)

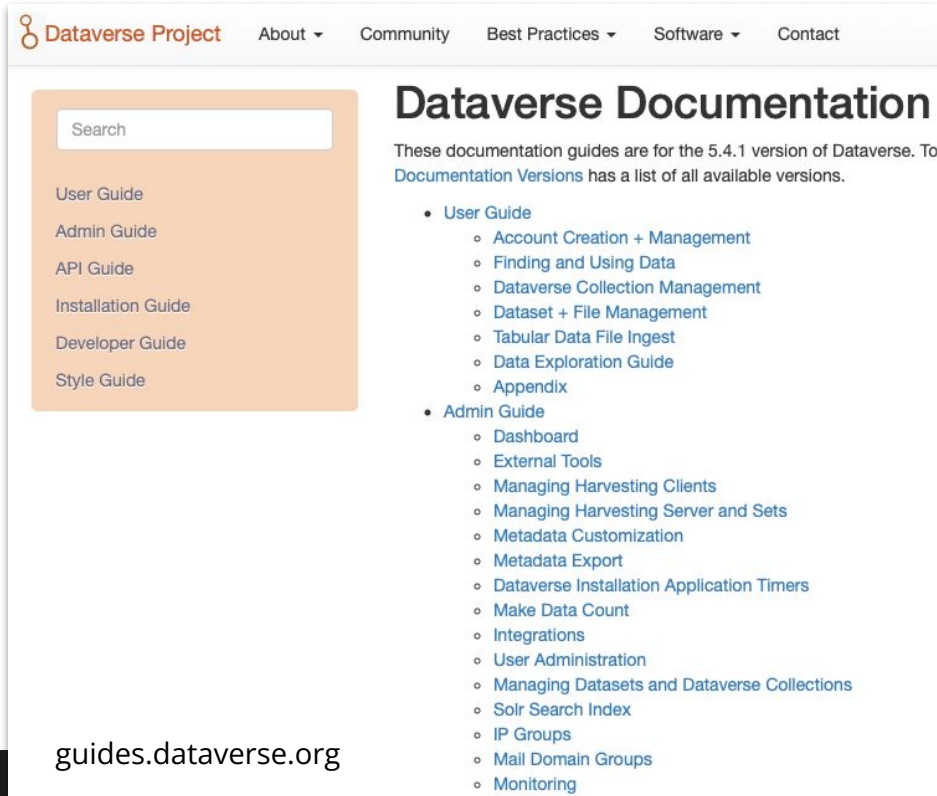
guides.dataverse.org

Documentation and guidelines

The Dataverse Project team maintains an extensive set of guidelines for repository managers, developers and users.

Dataverse Community meetings

- projects.iq.harvard.edu/dcm2021
- June 15-17, 2021



The screenshot shows the Dataverse Project documentation website. The header includes the Dataverse Project logo and navigation links: About, Community, Best Practices, Software, and Contact. The main content area is titled "Dataverse Documentation" and includes a search bar and a list of documentation guides. The guides are categorized into User Guide, Admin Guide, and API Guide, each with a list of sub-topics.

Dataverse Project About Community Best Practices Software Contact

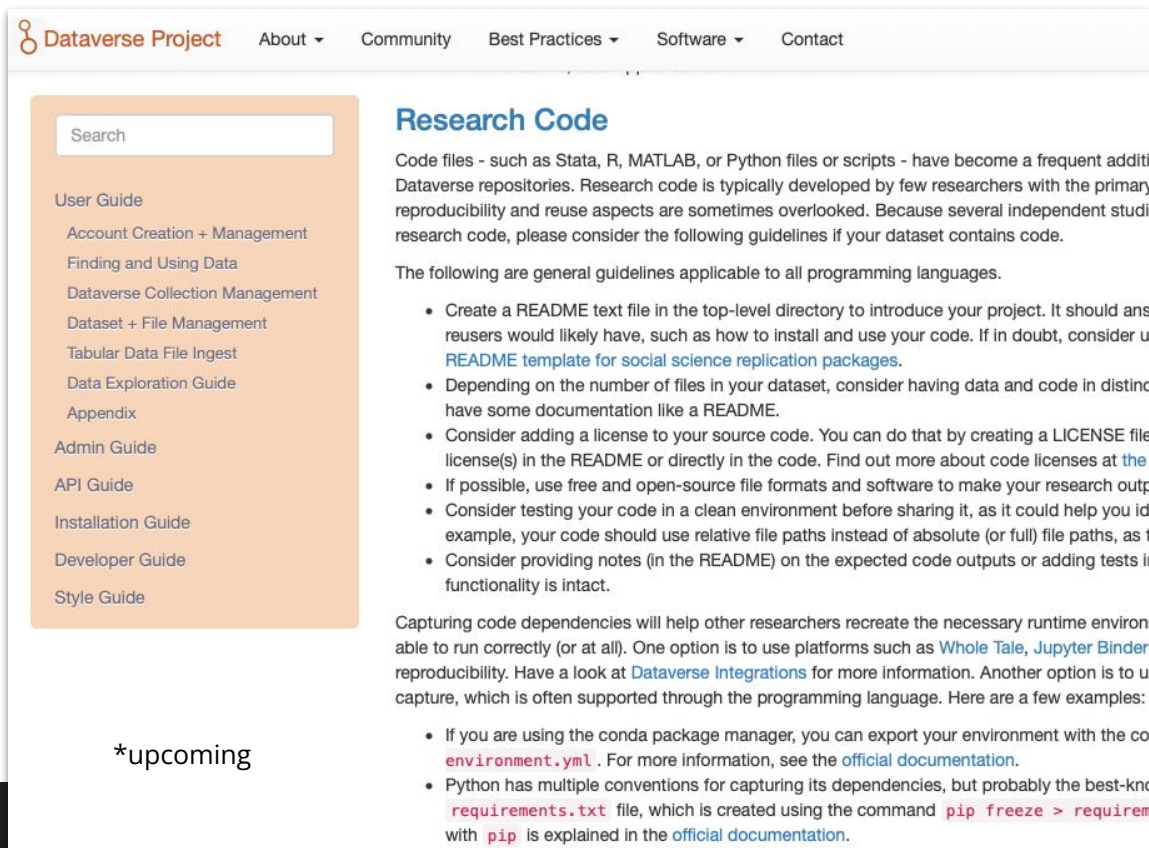
Dataverse Documentation

These documentation guides are for the 5.4.1 version of Dataverse. To [Documentation Versions](#) has a list of all available versions.

- **User Guide**
 - [Account Creation + Management](#)
 - [Finding and Using Data](#)
 - [Dataverse Collection Management](#)
 - [Dataset + File Management](#)
 - [Tabular Data File Ingest](#)
 - [Data Exploration Guide](#)
 - [Appendix](#)
- **Admin Guide**
 - [Dashboard](#)
 - [External Tools](#)
 - [Managing Harvesting Clients](#)
 - [Managing Harvesting Server and Sets](#)
 - [Metadata Customization](#)
 - [Metadata Export](#)
 - [Dataverse Installation Application Timers](#)
 - [Make Data Count](#)
 - [Integrations](#)
 - [User Administration](#)
 - [Managing Datasets and Dataverse Collections](#)
 - [Solr Search Index](#)
 - [IP Groups](#)
 - [Mail Domain Groups](#)
 - [Monitoring](#)

guides.dataverse.org

New guidelines



The screenshot shows the Dataverse Project website. The navigation bar includes 'Dataverse Project', 'About', 'Community', 'Best Practices', 'Software', and 'Contact'. A search bar is located in the top left. A sidebar on the left contains a list of guides: User Guide, Account Creation + Management, Finding and Using Data, Dataverse Collection Management, Dataset + File Management, Tabular Data File Ingest, Data Exploration Guide, Appendix, Admin Guide, API Guide, Installation Guide, Developer Guide, and Style Guide. The main content area is titled 'Research Code' and contains the following text and list:

Research Code

Code files - such as Stata, R, MATLAB, or Python files or scripts - have become a frequent addition to Dataverse repositories. Research code is typically developed by few researchers with the primary focus on reproducibility and reuse aspects are sometimes overlooked. Because several independent studies have shown that research code, please consider the following guidelines if your dataset contains code.

The following are general guidelines applicable to all programming languages.

- Create a README text file in the top-level directory to introduce your project. It should answer questions that users would likely have, such as how to install and use your code. If in doubt, consider using the [README template for social science replication packages](#).
- Depending on the number of files in your dataset, consider having data and code in distinct directories and have some documentation like a README.
- Consider adding a license to your source code. You can do that by creating a LICENSE file or license(s) in the README or directly in the code. Find out more about code licenses at the [Creative Commons website](#).
- If possible, use free and open-source file formats and software to make your research output more accessible.
- Consider testing your code in a clean environment before sharing it, as it could help you identify issues. For example, your code should use relative file paths instead of absolute (or full) file paths, as this makes it easier to run in a different environment.
- Consider providing notes (in the README) on the expected code outputs or adding tests if applicable to ensure functionality is intact.

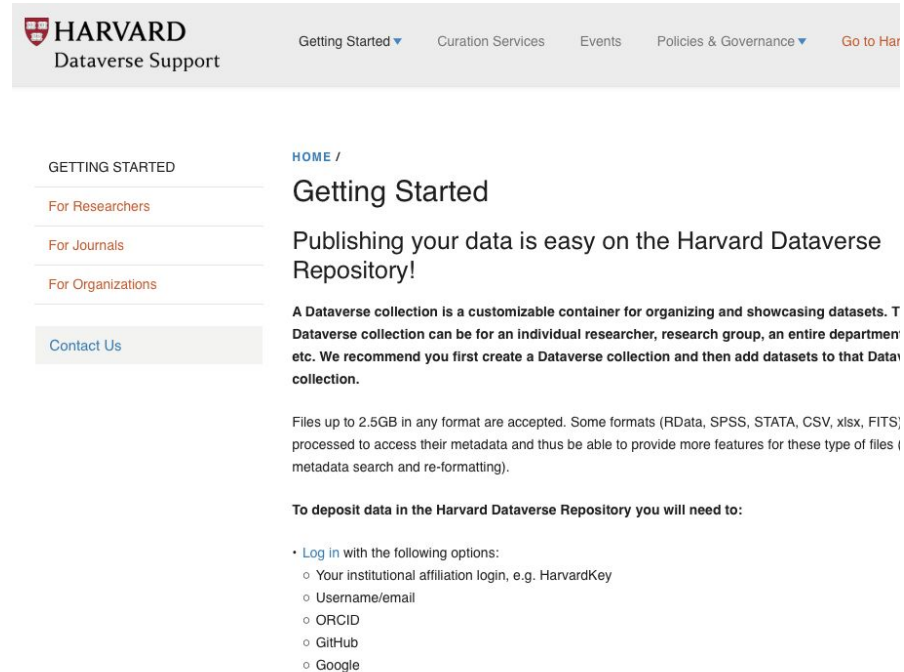
Capturing code dependencies will help other researchers recreate the necessary runtime environment to run correctly (or at all). One option is to use platforms such as [Whole Tale](#), [Jupyter Binder](#), or [Docker](#) for reproducibility. Have a look at [Dataverse Integrations](#) for more information. Another option is to use `conda` to capture dependencies, which is often supported through the programming language. Here are a few examples:

- If you are using the `conda` package manager, you can export your environment with the command `conda env export --name <environment> --no-builds > environment.yml`. For more information, see the [official documentation](#).
- Python has multiple conventions for capturing its dependencies, but probably the best-known is `requirements.txt`, which is created using the command `pip freeze > requirements.txt`. The `requirements.txt` file, which is created using the command `pip freeze > requirements.txt` with `pip` is explained in the [official documentation](#).

*upcoming

Documentation and guidelines

- Each Dataverse installation may have local documentation and support
- At Harvard, user training is organized by the Harvard Library and IQSS Data Curation Services



The screenshot shows the Harvard Dataverse Support website. The header includes the Harvard logo and 'Dataverse Support' on the left, and navigation links for 'Getting Started', 'Curation Services', 'Events', 'Policies & Governance', and 'Go to Har' on the right. The main content area is titled 'GETTING STARTED' and features a sidebar with links for 'For Researchers', 'For Journals', 'For Organizations', and 'Contact Us'. The main text area is titled 'Getting Started' and contains the following content:

HOME / Getting Started

Publishing your data is easy on the Harvard Dataverse Repository!

A Dataverse collection is a customizable container for organizing and showcasing datasets. T Dataverse collection can be for an individual researcher, research group, an entire department etc. We recommend you first create a Dataverse collection and then add datasets to that Data collection.

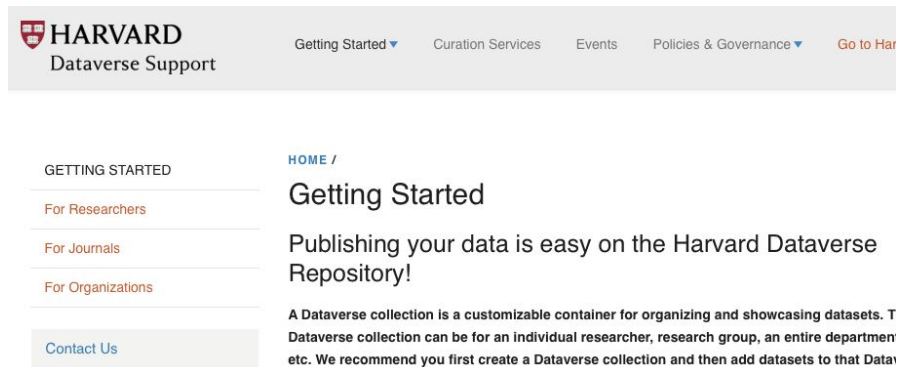
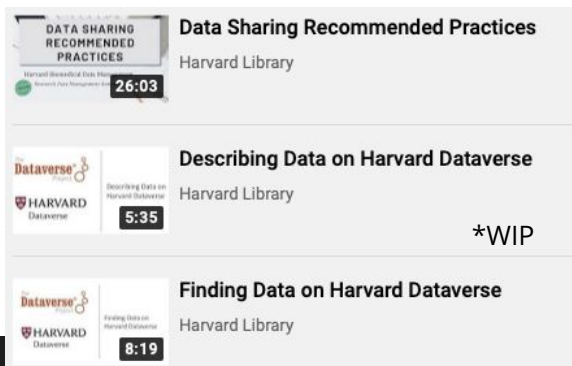
Files up to 2.5GB in any format are accepted. Some formats (RData, SPSS, STATA, CSV, xlsx, FITS) processed to access their metadata and thus be able to provide more features for these type of files (metadata search and re-formatting).

To deposit data in the Harvard Dataverse Repository you will need to:

- **Log in** with the following options:
 - Your institutional affiliation login, e.g. HarvardKey
 - Username/email
 - ORCID
 - GitHub
 - Google

Documentation and guidelines

- Each Dataverse installation may have local documentation and support
- At Harvard, user training is organized by the Harvard Library and IQSS Data Curation Services
- Video guides on YouTube



Documentation and guidelines

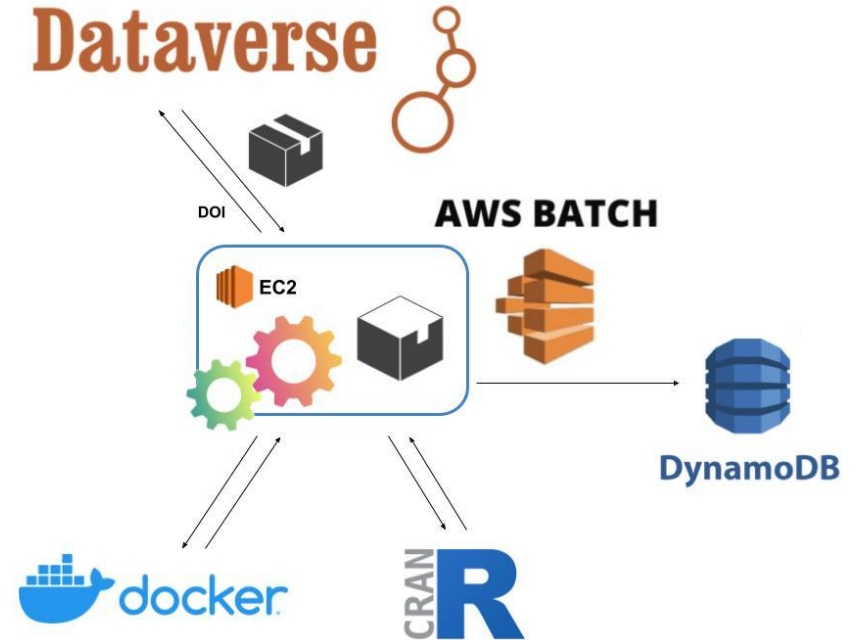
Summary

- Documentation and guidelines (global) are actively maintained
 - Institutional guidelines, support and training
-

Reproducibility and reuse

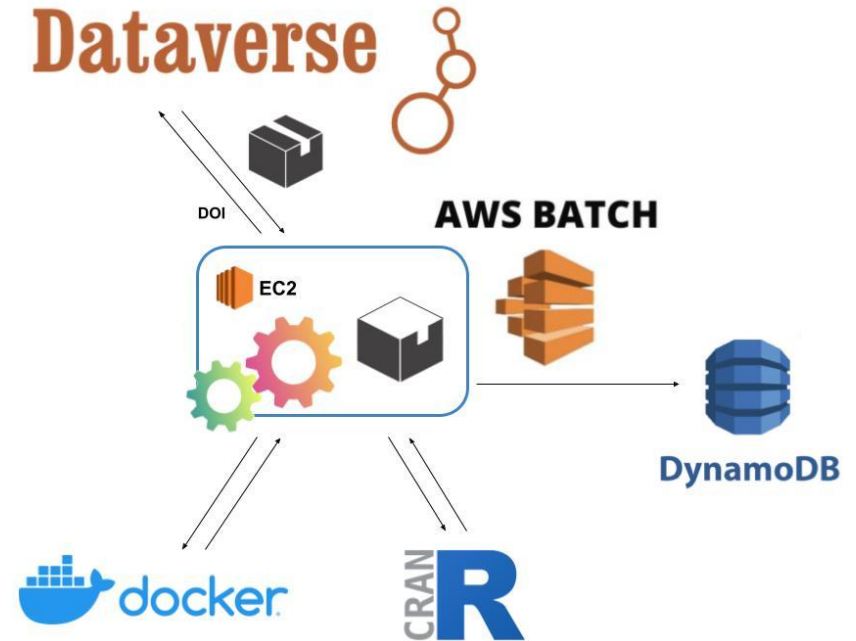
Reproducibility and reuse

- Reproducibility: “obtaining consistent computational results using the same input data, steps, code, and conditions of analysis”
- Conducted large scale analysis to examine execution and quality of R and Python code
- We retrieved replication datasets from Harvard Dataverse repository and R code was re-executed in a pre-installed Docker container

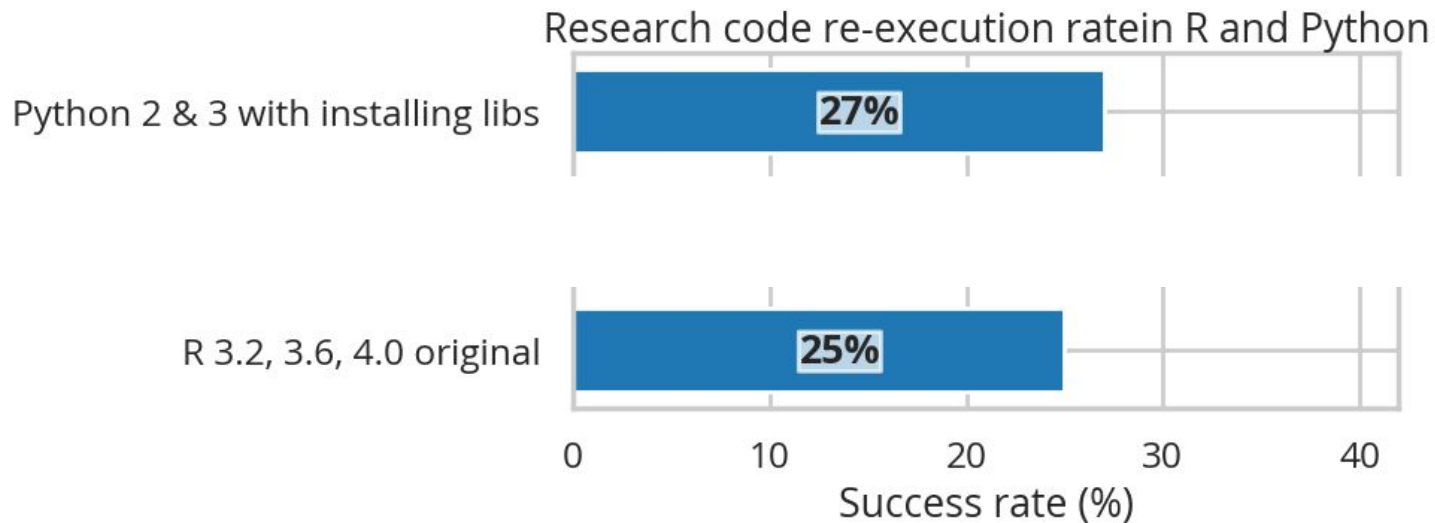


Reproducibility and reuse

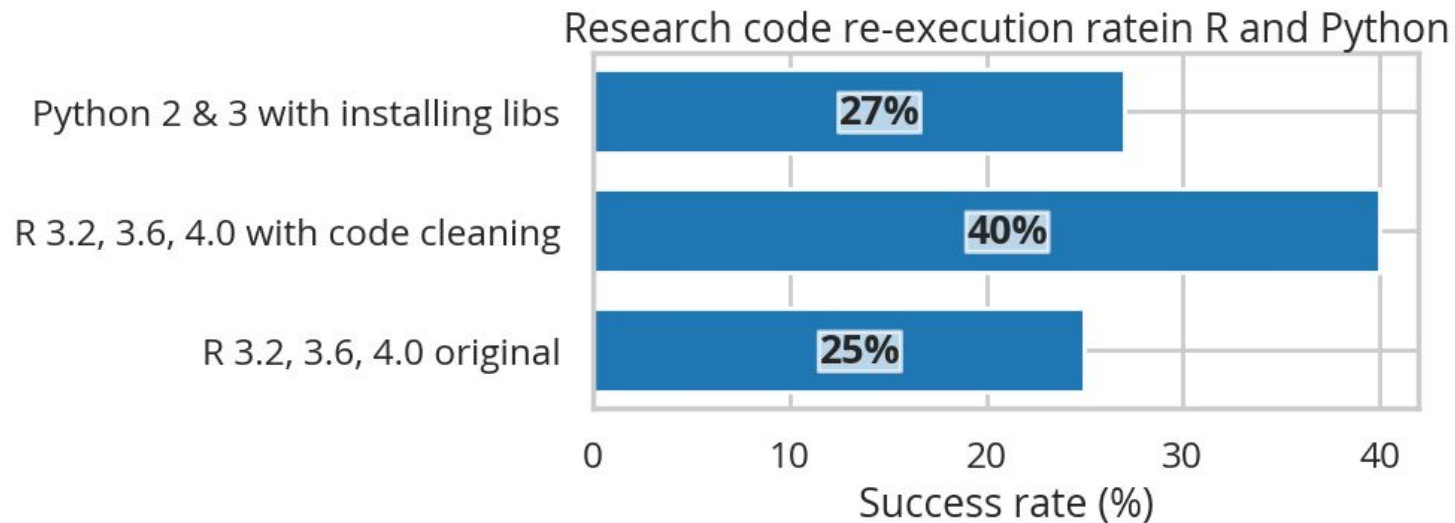
- Reproducibility: “obtaining consistent computational results using the same input data, steps, code, and conditions of analysis”
- Conducted large scale analysis to examine execution and quality of R and Python code
- We retrieved replication datasets from Harvard Dataverse repository and R code was re-executed in a pre-installed Docker container



Reproducibility and reuse

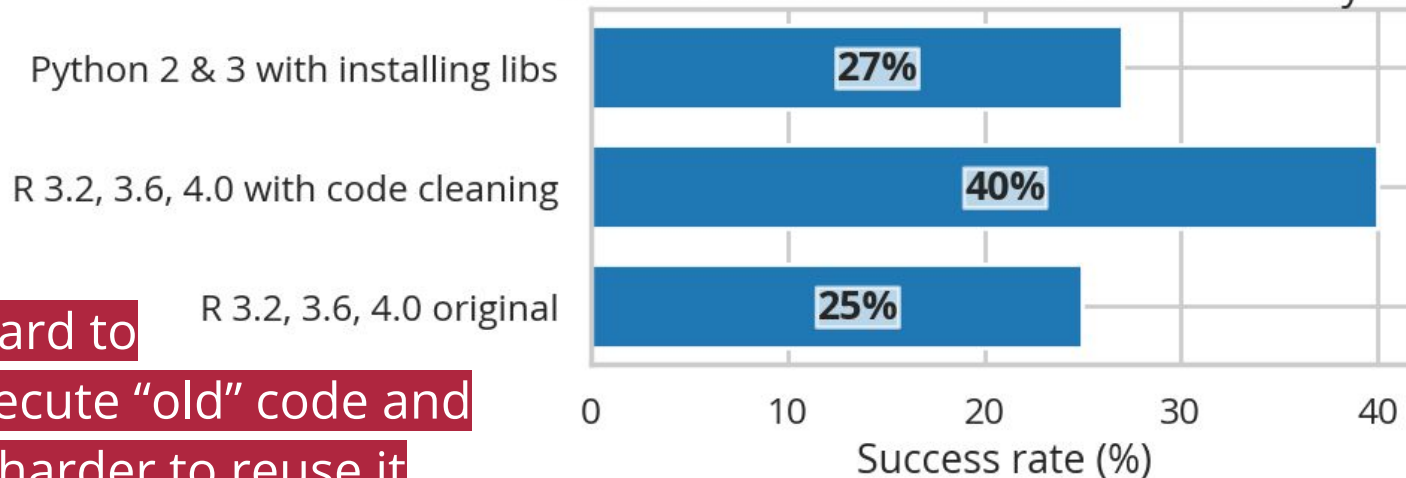


Reproducibility and reuse



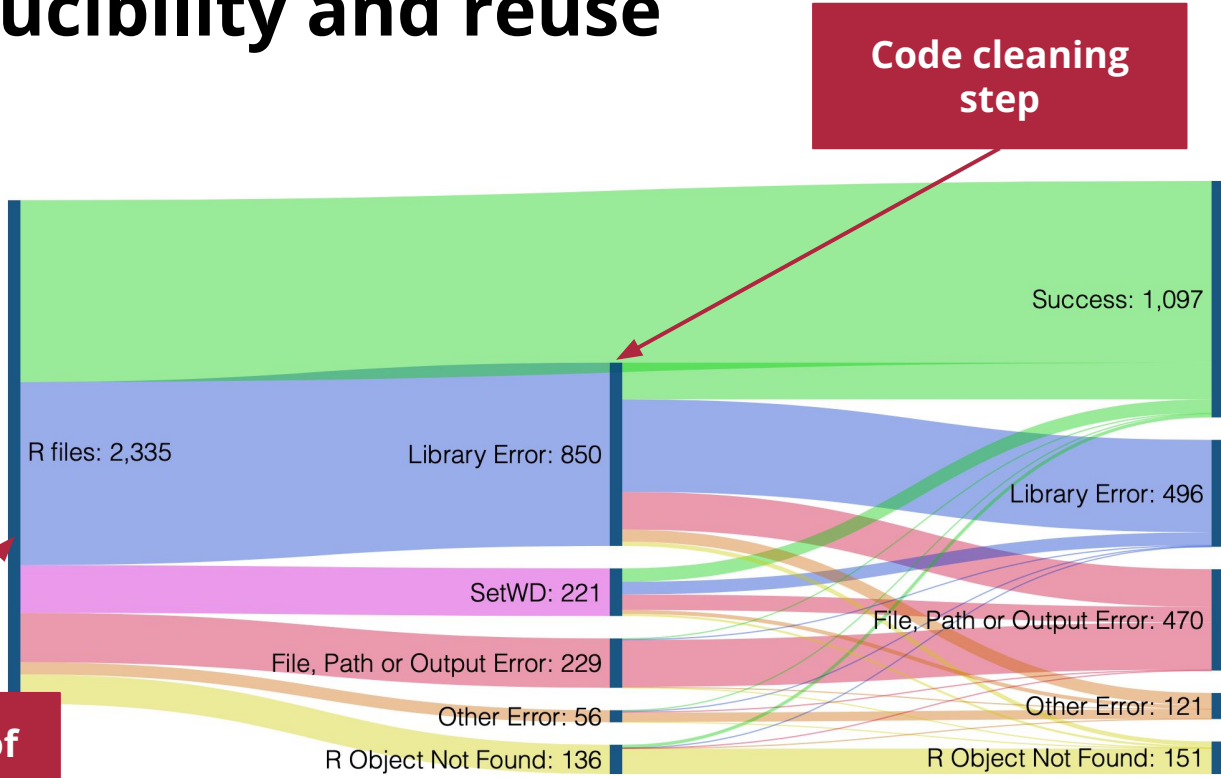
Reproducibility and reuse

Research code re-execution rate in R and Python



It is hard to re-execute "old" code and even harder to reuse it

Reproducibility and reuse

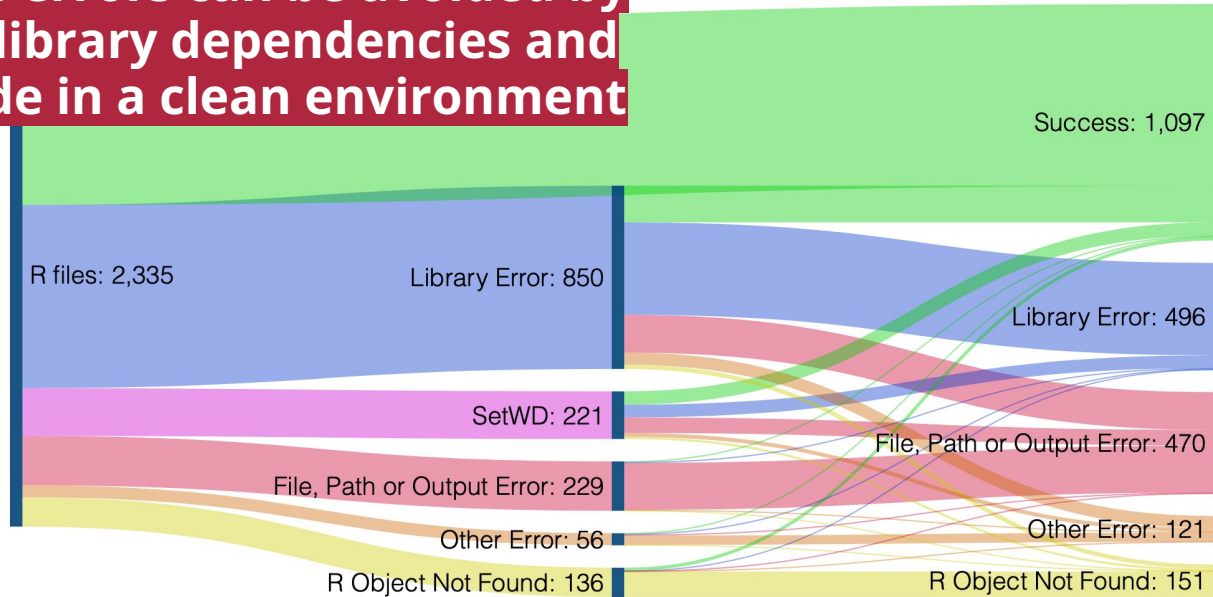


Re-execution of original code

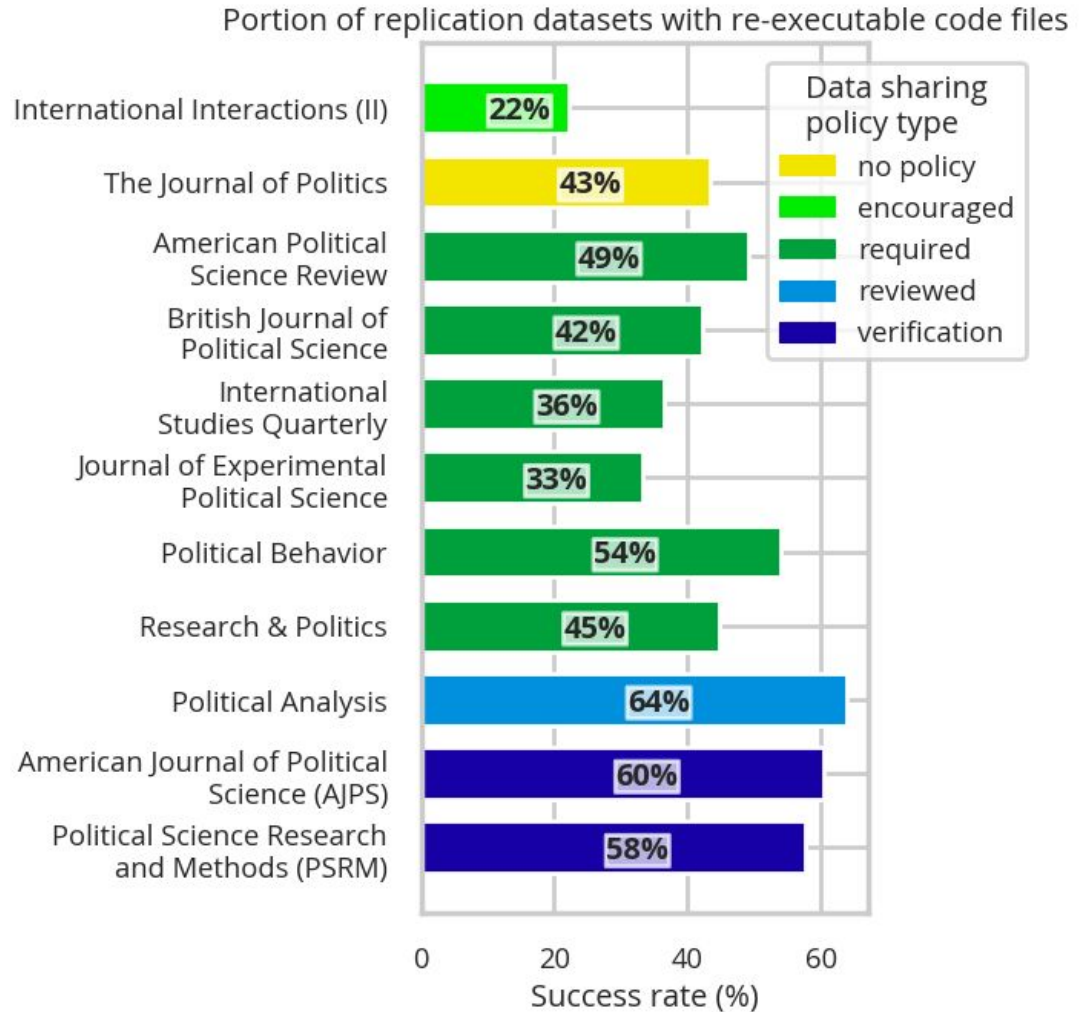
Code cleaning step

Reproducibility and reuse

Many code errors can be avoided by capturing library dependencies and testing code in a clean environment



Journals with stricter data policies have higher rate of executable code



Integration with reproducibility platforms

- New cloud tools have emerged to support collaborative work and research reproducibility by capturing code dependencies inside a web browser
- Dataverse software integration with reproducibility platforms allows:
 - Importing new research data with necessary libraries
 - Exporting and reusing the existing data



CODE OCEAN



In-browser view of the CodeOcean platform

The screenshot displays the CodeOcean interface for a project titled "Combo Synergy of Two Drugs". The main editor shows R code for data retrieval and visualization. The code includes the following key sections:

```
1 require(RSQLite)
2 require(plotly)
3
4
5 make_plot<-function(drugA, drugB, cell){
6   query<-paste('select drugA_name, drugA_conc, drugB_name, drugB_conc, x2x8, HSA, Bliss from c
7   drugA, "", drugB, ""') and drugB_name in ('', drugA, "", drugB, "") and cell_line=""
8   cell, '' order by drugA_conc, drugB_conc', sep='')
9
10  data<-dbGetQuery(db,query)
11
12  if (dim(data)[1]>0) {
13    drugA_conc<-unique(data[, "drugA_conc"])
14    drugB_conc<-unique(data[, "drugB_conc"])
15    a_conc<-rep(c(0,1,2,3),each=4)
16    b_conc<-rep(c(0,1,2,3),4)
17
18    observed.data<-matrix(data[, "x2x8"], nrow=4, ncol=4, byrow=TRUE)
19    hsa<-matrix(data[, "HSA"], nrow=4, ncol=4, byrow=TRUE)
20    bliss<-matrix(data[, "Bliss"], nrow=4, ncol=4, byrow=TRUE)
21
22    p<-plot_ly(z=observed.data, type="surface")
23    add_trace(x=b.conc, y=a.conc, z=observed.data, type="scatter3d", mode="points")
24    add_trace(x=b.conc, y=a.conc, z=hsa, type="surface")
25    add_trace(x=b.conc, y=a.conc, z=bliss, type="surface")
26    add_trace(x=b.conc, y=a.conc, z=bliss, type="scatter3d", mode="points")
27
28    layout(title = paste(drugA, drugB, cell),
29           scene = list(
30             zaxis = list(title = "viability"),
31             xaxis = list(title = "drugA concentration"),
32             yaxis = list(title = "drugB concentration"),
33             aspectratio=list(x=1, y=1, z=0.5)
34           )
35         )
36    htmlwidgets::saveWidget(as.widget(p), "Figure1.png")
37  }
38  else {
39    write(" Sorry, This combination does not exist")
40  }
41 }
```

The right-hand side of the interface shows a "Results" panel with a list of runs:

Run ID	Status
Run 90665	Completed
Run 9660725	Completed
Run 9655826	Completed
Run 9650940	Completed
Run 3563692	Completed
Run 3363841	Completed
Run 2826088	Completed
Published Result	Published

The "Input Files" section shows a file named "DrugCombo.db" with a size of 115.76 MB.

The 3D plot, titled "Figure1.png" and "5-FU & ABT-888 Combination", visualizes the synergy between the two drugs. The vertical axis represents "viability" (ranging from 0.6 to 1.0), the horizontal axis represents "drugA concentration" (ranging from 0.35 to 1.08), and the depth axis represents "drugB concentration" (ranging from 0.35 to 1.08). The plot features three surfaces: a blue surface for "Observed Data", a red surface for "Predicted (HSA)", and a green surface for "Predicted (Bliss)". Data points are shown as colored dots corresponding to these surfaces. The plot shows a complex, multi-faceted surface with several peaks and valleys, indicating non-linear synergy between the drugs.

Integration with reproducibility platforms

- New cloud tools have emerged to support collaborative work and research reproducibility by capturing code dependencies inside a web browser
- Dataverse software integration with reproducibility platforms allows:
 - Importing new research data with necessary libraries
 - Exporting and reusing the existing data



CODE OCEAN



Example: Jupyter Binder and Harvard Dataverse repository

Replication Data for: Repository approaches to improving quality of shared data and code

Version 4.1



Trisovic, Ana, 2020, "Replication Data for: Repository approaches to improving quality of shared data and code", <https://doi.org/10.7910/DVN/EA3LC5>, Harvard Dataverse, V4

Cite Dataset Learn about Data Citation Standards.

Access Dataset

- Edit Dataset
- Link Dataset
- Contact Owner
- Share



Description

This is supplementary data to the article "Repository approaches to improving quality of shared data and code," and in particular, its first section on complete research code.

Run this code on Jupyter Binder here: [launch binder](#) (2020-09-27)

Subject

Computer and Information Science

Files Metadata Terms Versions

Search this dataset...

Find

Upload Files

Example: Jupyter Binder and Harvard Dataverse repository



Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

New to Binder? Get started with a Zero-to-Binder tutorial in [Julia](#), [Python](#) or [R](#).

Build and launch a repository

Dataverse DOI (10.7910/DVN/TJCLKP)

Dataverse DOI (10.7910/DVN/TJCLKP)

Git ref (branch, tag, or commit)

HEAD

Path to a notebook file (optional)

Path to a notebook file (optional)

mybinder.org

Replication Data for: Repository approaches to improving quality of shared data and code

Version 4.1



Trisovic, Ana, 2020, "Replication Data for: Repository approaches to improving quality of shared data and code", <https://doi.org/10.7910/DVN/EA3LC5>, Harvard Dataverse, V4

Cite Dataset Learn about Data Citation Standards.

Access Data
Edit Data
Link Data
Contact Owner

Dataset Metrics

22 Downloads

Description

This is supplementary data to the article "Repository approaches to improving quality of shared data and code," and in particular, its first section on completeness of research code.

Run this code on Jupyter Binder here: [launch binder](#) (2020-09-27)

Subject

Computer and Information Science

Files Metadata Terms Versions

Search this dataset...

Find

+ Upload Files

Example: Jupyter Binder and Harvard Dataverse repository

Build and launch a repository

Dataverse DOI (10.7910/DVN/TICLKP)

Branch, tag, or commit) Path to a notebook file (optional)

Path to a notebook file (optional) File launch

Copy the URL below and share your Binder with others:

<https://mybinder.org/v2/dataverse/10.7910/DVN/EA3LC5/>

Expand to see the text below, paste it into your README to show a binder badge: [launch binder](#)

Waiting Building

Build logs hide

```
Found existing installation: Jinja2 2.11.3
INFO:repo2docker: Found existing installation: Jinja2 2.11.3

Uninstalling Jinja2-2.11.3:
INFO:repo2docker: Uninstalling Jinja2-2.11.3:

Successfully uninstalled Jinja2-2.11.3
INFO:repo2docker: Successfully uninstalled Jinja2-2.11.3

Attempting uninstall: ipython
INFO:repo2docker: Attempting uninstall: ipython

Found existing installation: ipython 7.20.0
INFO:repo2docker: Found existing installation: ipython 7.20.0
```

mybinder.org

Jupyter Binder
creates runtime
environment for
research code

Example: Jupyter Binder and Harvard Dataverse repository

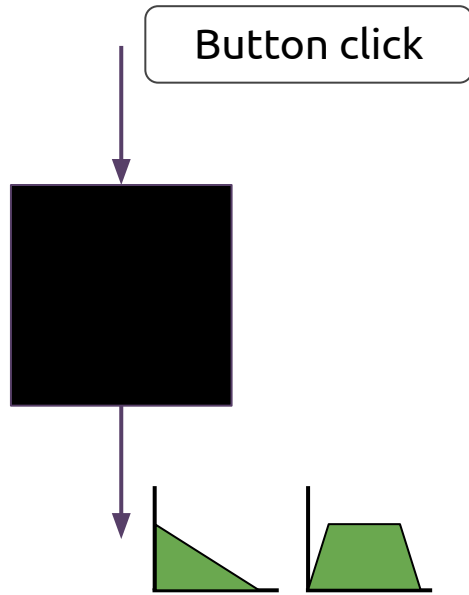
Anyone can explore data, make direct changes in the code or data

The screenshot shows a Jupyter Binder interface with the following components:

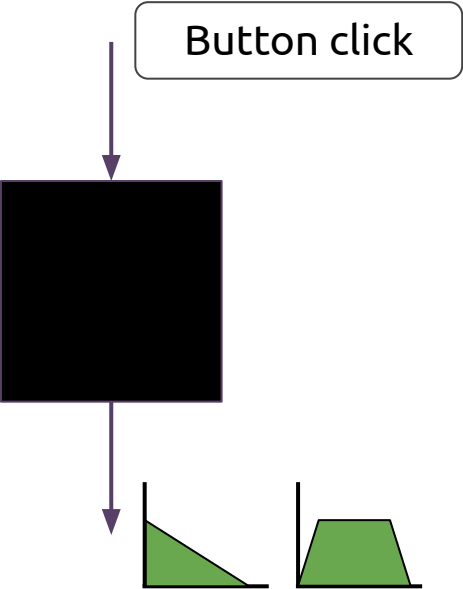
- Browser Header:** hub-binder.mybinder.ovh, Start Page, Visit repo, Copy Binder link, Python 3, Memory: 203.7 MB / 2 GB.
- Code Editor:** Analysis of Python results. The code includes imports for numpy, matplotlib, seaborn, pandas, and plot styles.
- Code Execution:** In [1]: imports; In [2]: df = pd.read_csv("python-study-data.csv", index_col=0); In [3]: df.head().
- Output Table:** A table with 6 columns: doi, filename, result2, result3, list_of_all, and size. It contains 5 rows of data, each with a different filename and associated errors in result2 and result3.
- Code Execution:** In [91]: df[df.result2.isnull()];
- Output Table:** A table with 6 columns: doi, filename, result2, result3, list_of_all, and size. It is currently empty.

	doi	filename	result2	result3	list_of_all	size
0	doi:10.7910/DVN/8TB7GO	ei_preprocessing.py	TypeError: coercing to Unicode: need string or...	TypeError: invalid file: None	simulation_output.txt;format_cdc_data.sh;evalu...	274991
1	doi:10.7910/DVN/8TB7GO	ei_preprocessing_india.py	TypeError: unsupported operand type(s) for +: ...	SyntaxError: Missing parentheses in call to 'p...	simulation_output.txt;format_cdc_data.sh;evalu...	274991
2	doi:10.7910/DVN/8TB7GO	ei_preprocessing_ipums_census_acs_samples.py	AttributeError: 'NoneType' object has no attri...	SyntaxError: Missing parentheses in call to 'p...	simulation_output.txt;format_cdc_data.sh;evalu...	274991
3	doi:10.7910/DVN/8TB7GO	ei_preprocessing_ipums_full_census.py	AttributeError: 'NoneType' object has no attri...	SyntaxError: Missing parentheses in call to 'p...	simulation_output.txt;format_cdc_data.sh;evalu...	274991
4	doi:10.7910/DVN/8TB7GO	ei_preprocessing_race.py	TypeError: coercing to Unicode: need string or...	TypeError: invalid file: None	simulation_output.txt;format_cdc_data.sh;evalu...	274991

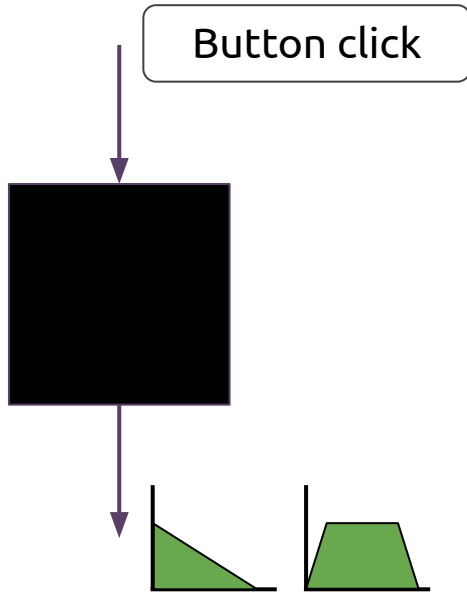
Reproducibility versus reuse



Reproducibility versus reuse



Reproducibility versus reuse



Harvard Dataverse > Murray Research Archive Dataverse > Early Head Start Research and Evaluation Project Dataverse >

Early Head Start Research and Evaluation Project, 1996 - 2001

Version 11.0

Administration for Children and Families, 2009, "Early Head Start Research and Evaluation Project, 1996 - 2001", <https://doi.org/10.7910/DVN/TH7GEB>, Harvard Dataverse, V11

[Cite Dataset](#) - [Learn about Data Citation Standards.](#)

[Access Dataset](#) - [Contact Owner](#) [Share](#)

Dataset Metrics [📄](#)
5,230 Downloads [📄](#)

Description [📄](#) This study page contains cataloging and documentation files (only) related to the *Early Head Start* data archived in the Murray Research Archive Dataverse.

The purpose of this study was to assess the impact of early head start programs in response to the 1994 Head Start reauthorization which established a special initiative for services to families with infants and toddlers. The study was a program evaluation with 1500 families in Early Head Start programs and 1500 in a control group with no program participation.

The participants included 30 00 low-income and poor families (child, mother, and some fathers). The [dataset page](#) contains a complete list of files, which can be filtered by file name, file type, access level, etc.

[Read full Description \[+\]](#)

Subject [📄](#) Social Sciences

Keyword [📄](#) Early Head Start, Education, Low Income, Poor Families, African-American, Fathers, Latino

Related Publication [📄](#) Friberg, B. L. (2010). Testing theoretical models of aggression and sustained attention development within the context of Early Head Start. (Doctoral dissertation). University of Wisconsin, Madison, WI.

[Files](#) [Metadata](#) [Terms](#) [Versions](#)

Search this dataset... [🔍](#)

Filter by
File Type: All - Access: All - File Type: All -

[📄 Sort](#)

1 to 10 of 47 Files		Download
<input type="checkbox"/>	00097_Early_Head_Start_Baseline (HSFIS) data files.doc MS Word - 62.5 KB Published Oct 13, 2013 54 Downloads MDS: 5da...09c 📄 User explanation for Baseline data 📄 Viewing page	📄
<input type="checkbox"/>	00097_Early_Head_Start_provider crosswalk_public_ju07.doc MS Word - 109.0 KB Published Oct 13, 2013 22 Downloads MDS: 4aa...c28 📄 Shows variable-level relationships between different center and family (Home) interviews, including initial, "repeat visit", and "annual update" versions 📄 Children's Headstart page	📄
<input type="checkbox"/>	00097_Early_Head_Start_PSI_x-walk public.doc MS Word - 448.5 KB Published Oct 13, 2013 28 Downloads MDS: a85...544 📄 Early Head Start Parent Services Interview (PSI) Cross Walk (revised June 2006) 📄 Parent Services Interview and Exit Interview	📄

Research and prototype projects

Research reproducibility test at
upload

Ana Trisovic

```
1 Project reference: sJV8vKMPYjutkCOBQmWR
2 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
3 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
4 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
5 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
6 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
7 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
8 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
9 FILENAME: ButlerHomola_Excludability_Appendix.R
10 RESULT: ["Error in library("stargazer") : there is no package called 'stargazer'"]
11 FILENAME: ButlerHomola_Excludability_Analysis.R
12 RESULT: ["Error in library("stargazer") : there is no package called 'stargazer'"]
13
```

Research and prototype projects

Research reproducibility test at upload

Code readability assessment

My Account Sign out

Ana Trisovic

```
1 Project reference: sJV8vKMPYjutcCOBQmWR
2 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
3 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
4 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
5 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
6 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
7 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
8 downloading file: https://firebasestorage.googleapis.com/v0/b/re3-fb.appspot.com/o/rep
9 FILENAME: ButlerHomola_Excludability_Appendix.R
10 RESULT: ["Error in library("stargazer") : there is no package called 'stargazer'"]
11 FILENAME: ButlerHomola_Excludability_Analysis.R
12 RESULT: ["Error in library("stargazer") : there is no package called 'stargazer'"]
13
```

My Account Sign out

Get Code Rating

6.34

```
15
16 ## read in "Names" dataset (Table 1 & Figure 1)
17 data <- read.csv("ButlerHomola_Excludability_Names.csv")
18
19 ## Remove name with no match (Alaliyah)
20 data <- data[-2,]
21
22 ## Create political resources factor score
23 resources <- prcomp(~ Income + Education + Housing1 + Turnout,
24                   data=data, na.action = na.exclude)
25 data$resources <- -1*(resources$x[,1])
26 ## Eigenvalue of first factor
27 (resources$sdev^2*4/sum(resources$sdev^2))[1]
28 ## 2.73
29
30 ## check that high factor values equal higher values on the component part
31 cor(data$Income, data$resources, use="complete.obs")
32 cor(data$Education, data$resources, use="complete.obs")
33 cor(data$Housing1, data$resources, use="complete.obs")
34 cor(data$Turnout, data$resources, use="complete.obs")
35
36 ## correlation matrix
37 cor(cbind(data$resources, data$Income, data$Education, data$Housing1,
38          data$Turnout), use="complete.obs", method="spearman")
39
```

Consider:

- breaking up your lines
- removing any irrelevant parentheses in your lines
- diminishing the periods in your lines

Upload File

Reproducibility and reuse

Summary

- A large-scale study sheds light on reproducibility challenges
 - Integrations with cloud tools facilitate software capture
 - Further development focuses on reuse
-

**Open questions,
research and
development**

Open questions, research and development

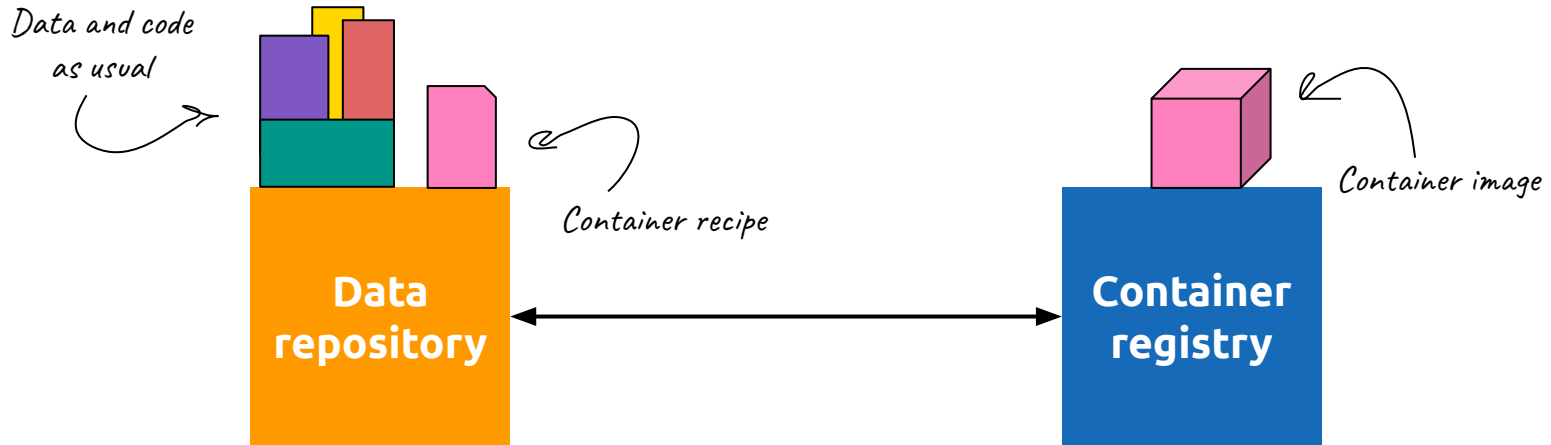
- DDI – Cross Domain Integration (DDI-CDI)
 - Planning to implement cross-domain metadata to facilitate merging data files
 - Controlled vocabularies for each variable
- Will incorporate software metadata
 - Enabling licencing, attribution and software dependency
- Global Dataverse Community Consortium (GDCC) - facilitates Dataverse community efforts, and supports Dataverse repositories around the world
 - GDCC working groups are formed to tackle the open questions

Open questions, research and development

- DDI – Cross Domain Integration (DDI-CDI)
 - Planning to implement cross-domain metadata to facilitate merging data files
 - Controlled vocabularies for each variable
- Will incorporate software metadata
 - Enabling licencing, attribution and software dependency
- Global Dataverse Community Consortium (GDCC) - facilitates Dataverse community efforts, and supports Dataverse repositories around the world
 - GDCC working groups are formed to tackle the open questions

Software, workflows and containers working group

- SWC (swc.gdcc.io) - identifies requirements and undertakes developments to support research software and reproducibility at the Dataverse Project
 - We welcome new members!
 - One solution for containers:



Conclusion

- The Dataverse Project creates **research data repository** software
- Ongoing support and ongoing enhancements for descriptive metadata for datasets and variable metadata for tabular files
- Research is conducted to:
 - examine common practices,
 - identify shortcomings of existing approaches and
 - new software developments
- Further development include supporting dissemination of emerging computational components, such as **workflows** and **containers**, with specific metadata, tools and infrastructure

Conclusion

- The Dataverse Project creates **research data repository** software
- Ongoing support and ongoing enhancements for descriptive metadata for datasets and variable metadata for tabular files
- Research is conducted to:
 - examine common practices,
 - identify shortcomings of existing approaches and
 - new software developments
- Further development include supporting dissemination of emerging computational components, such as **workflows** and **containers**, with specific metadata, tools and infrastructure



ALFRED P. SLOAN
FOUNDATION



Thank you!

Email: anatrisovic@g.harvard.edu
GitHub & Twitter: [atrisovic](#)
Dataverse Project: <https://dataverse.org/contact>